CLUWLL Project Proposal

Marta Berardi, Pia Schwarz

January 23, 2020

1 Purpose

Our system consists of a vocabulary learning web application for L1 speakers of German willing to practice their knowledge of Italian as L2. The main goal of our app is to support vocabulary learning and reinforcement through context sentences and word focus with Fill-in-the-blank exercises (FIBs).

We believe that this configuration will facilitate the learning process - making it varied and therefore less boring for the user - and will promote incidental training of various forms of the lexical units.

2 Coverage

The learning material provided by our app will consist of 50 to 100 Italian texts, half of them of level B1 and the other half of level B2. The texts will therefore include words from the lexical lists of levels B1 and B2 (around 2000 words) outlined by the CEFR (cf. section 8.3).

Each text, together with its *vocabulary practice* and a *wrap-up* section will constitute a unit (cf. section 4, *Mechanics*, for more detailed information about the structure of the unit).

3 Target Group

We imagine our typical user to be a L1 speaker of German, studying Italian as L2 at a level ranging from B1 to B2. We believe our app to be suitable for learners of different ages, but not younger than 15/16 years old.
We also imagine our system to be the most useful if used as an additional tool for vocabulary learning and practice, either outside the classroom or during leisure time. Since our app does not provide any theoretical information about grammar, we believe it to yield the best results if combined with independent learning of grammatical rules - either in class or through the consultation of a different educational resource.

4 Mechanics

Each **practice unit** will consist of four blocks:

1. **Reading**: user reads a short text and clicks on all unknown words (target words)
2. **Vocabulary practice**:
    - per target word up to 3 context sentences are presented to grasp the meaning through context. 3-4 target word contexts are presented as a block. The target word can appear in different forms in the context sentences. The user has the possibility to check the translation of the sentences excluding the target word;
    - all context sentences from the previous block (9-12 sentences) are presented in a scrambled order as FIB exercises masking the target words. The 3-4 target word's lemmas are displayed at the same time as a list. The user gets simple feedback (correct/incorrect) and has two trials. For answers that are still incorrect after the second trial (= difficult words), the user has the option to check the translation of the target word;
    - the next block of context sentences + FIBs is provided until all target words are covered.
3. **Wrap up**: The same text from step 1 will be presented and the user marks again all still unknown words in the text (= difficult words). Then, information about the progress (before-after) will be displayed and the difficult words will be stored in the learner model.

Apart from the practice unit we want to offer **4 piles of vocabulary flashcards** to practice (they are available to the user anytime before and after the practice units):

- Bootstrap pile: Essential n words (presented with translation) that are necessary to start covering new words from level B1/B2, the app presupposes the knowledge of these words
- All learned words: a pile containing all words the user has encountered anywhere in the app
- Most difficult words: n most difficult words for the user
- Most frequent Italian words: n most frequent Italian words from the pile of all learned words

5 Learner Modeling

The learner model will keep track of:

- the words the user encountered so far during the usage of the app
- the words the user has difficulties with

6 User-adaptivity

Based on the learner model data we will provide suitable content to fill the flashcard piles.

7 Gamification

We will not use gamification to a big extent, the only part that can be seen as gamified is the practice unit's Wrap up (see section "Mechanics" - vocabulary practice).

8 Resources and NLP

8.1 Text retrieval

We will use didactic texts either retrievable from the Web or from old didactic material; another possible solution would be to consult books or corpora which are accessible in a computational way and which are available under a compatible license.

8.2 Translations

We will retrieve:

- translations of context sentences from Tatoeba;
- sentence alignments and translations through a Java library which will be provided to us by Dr. Dellert;
- if needed, we will use translations retrieved from a dictionary provided to us by Dr. Dellert, or from Wiktionary (https://www. wiktionary.org) from a previous project by Alina Baranova and Marta Berardi (https://github.com/AlinaBaranova/LearnIt/tree/master/Tables).

8.3 Level-specific lexical lists

The lexical lists of levels B1 and B2 will be retrieved from the Italian version of the Common European Framework of Reference for Languages, in short CEFR (Spinelli, B., Parizzi, F., 2011. Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2. Ed. La Nuova Italia.)

8.4 Inflection

We will make use of the code written by Marta Berardi in a previous project to derive inflections of verbs, nouns and adjectives (https://github.com/AlinaBaranova/LearnIt/tree/master/InflectionRules).

8.5 Context sentences

In order to retrieve context sentences to be employed in the vocabulary practice section, we will query Tatoeba (https://tatoeba.org) and, if needed, OpenSubtitles (http://opus.nlpl.eu/OpenSubtitles-v2018.php).

We need to keep in mind that a potential problem could arise while consulting the Tatoeba corpus, namely that some sentences might not be useful to understand the meaning of the word in context. We will explore the possibility of searching for sentences which contain at least 3 content words (= low frequency words) and whether it is possible to simplify them by lexical replacement, in case the content words are of a level above B2. For this, we might need to use an Italian WordNet (http://multiwordnet.fbk.eu/online/multiwordnet.php).
In case we can extract more than the 3 necessary high quality context sentences, we try to prioritize sentences that contain difficult words for the user and were also displayed n times to them.
In order to pre-process the text we will need to tokenize, lemmatize and POS-tag them. We intend to use the available TreeTagger tool for this purpose.

9 Nice-to-have Features

- vocabulary FIBs: more precise feedback through distinguishing wrong word errors and typos.